



技术：应用场景下的 数据基本分析流程和分析方法

夏菁

浙江大学CAD&CG国家重点实验室
可视化与可视分析小组



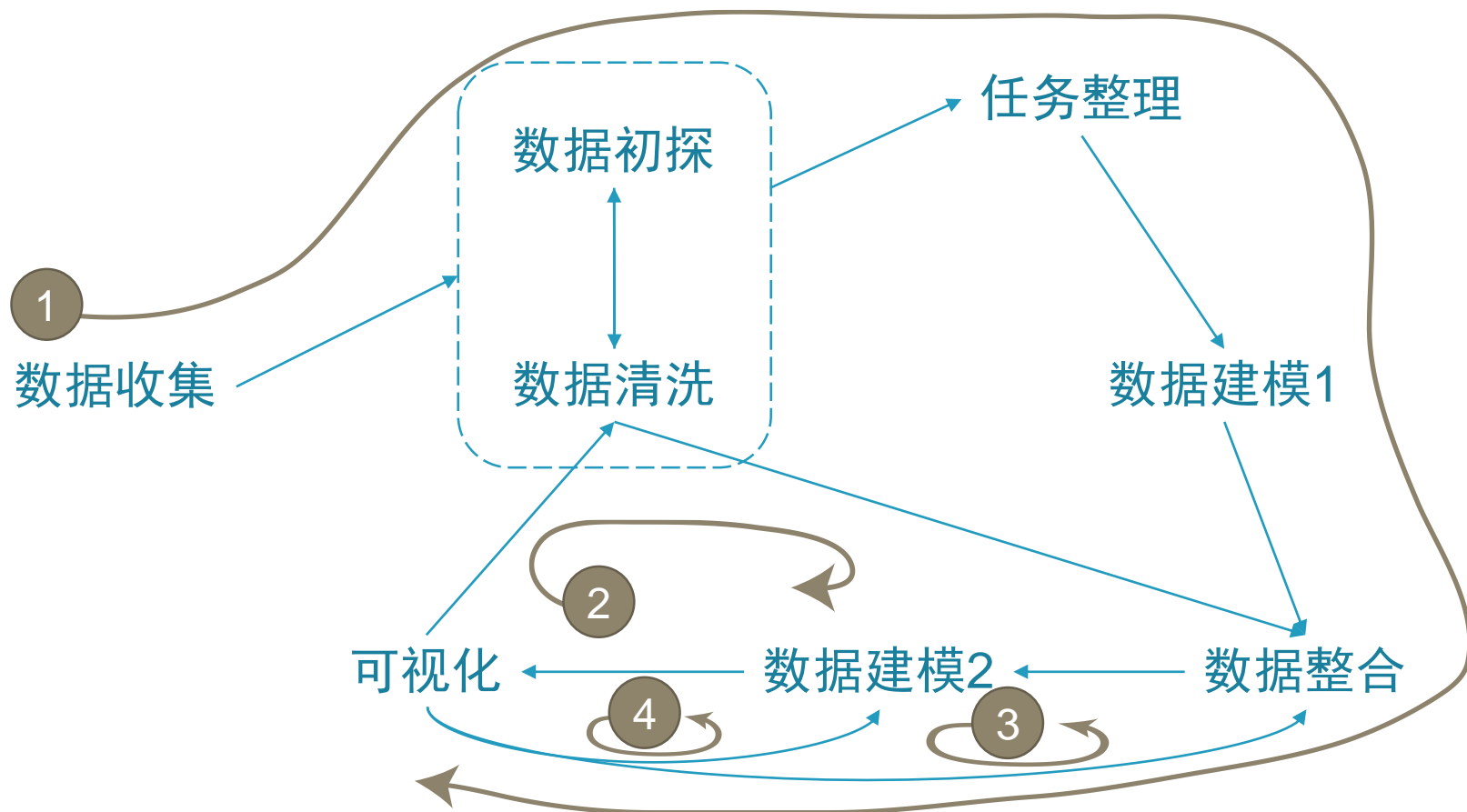
基础 1

数据属性

也可以称为变量、维度、特征等



数据分析流程



公共数据来源

名称	网址	备注
数据堂	http://www.datatang.com	国内科研数据平台
政府公开数据	http://www.data.gov	包括美国、印度、英国等各国政府公开数据
可视分析竞赛数据	http://www.vacommunity.org/tiki-index.php	历年可视分析会议的竞赛数据
全球恐怖主义行动数据	http://www.start.umd.edu/gtd	全球历年的恐怖主义事件统计
地图数据	Openstreetmap, google map, bing map, 高德地图,	开放地图API
社交数据	微博数据、Twitter数据等	社交网络的爬虫、API
维基日志数据	https://en.wikipedia.org/wiki/Wikipedia:Database_download	维基百科数据集合辑

数据质量

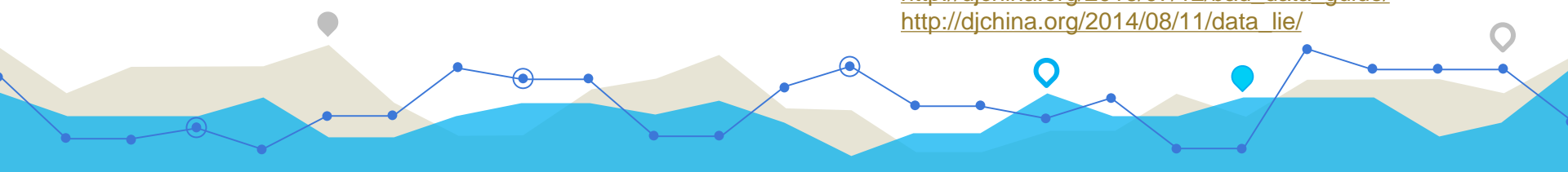
V	有效性	数据是否真实合理
A	准确性	数据是否精确，有无误差
B	可信性	数据来源和收集方式是否可信
I	一致性	数据(格式、单位等)是否一致
C	完整性	数据是否有缺失
T	时效性	数据适用范围(相对分析任务)

Kandel S, Heer J, Plaisant C, et al. Research directions in data wrangling: Visualizations and transformations for usable and credible data[J]. Information Visualization, 2011, 10(4): 271-288.

The Quartz 坏数据手册

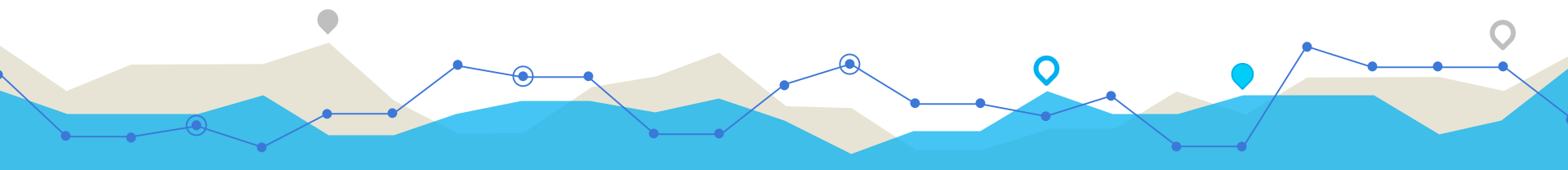
http://djchina.org/2016/07/12/bad_data_guide/

http://djchina.org/2014/08/11/data_lie/



数据清洗

- 确定数据属性类型
- 确定数据在各属性的值分布，纠正不合理值
- 纠正数据的不一致性
- 校准数据



城市数据质量



乒乓效应：移动通信系统中，如果在一定区域里两基站信号强度剧烈变化，手机就会在两个基站间来回切换，产生所谓的“乒乓效应”。



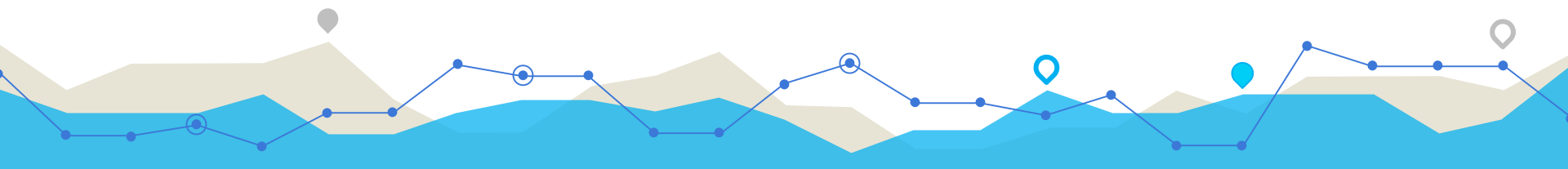
数据噪声：简单来说，数据噪声指在一组数据中无法解释的数据变动，就是一些不和其他数据相一致的数据。



数据缺失：有些数据点的数据存在数据值为空的情况，可能导致分析效果出现错误及偏差



数据一致性：数据存储的一致性模型可以认为是存储系统和数据使用者之间的一种约定。



城市数据的清洗



数据去噪

主要处理数据噪声，去掉无意义的数
据。并对错误数据进行检测和去除



数据补全：

对缺失数据进行补全

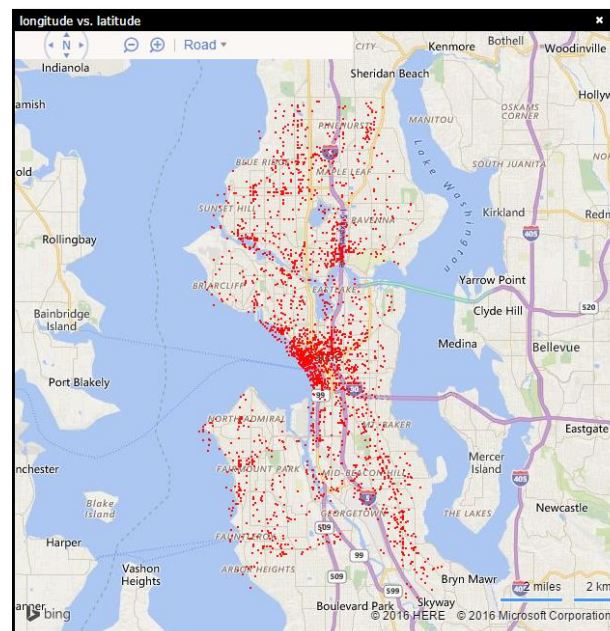
数据更正：

对数据进行数据修改，包括处理乒乓
效应，统一数据格式和数据一致性。



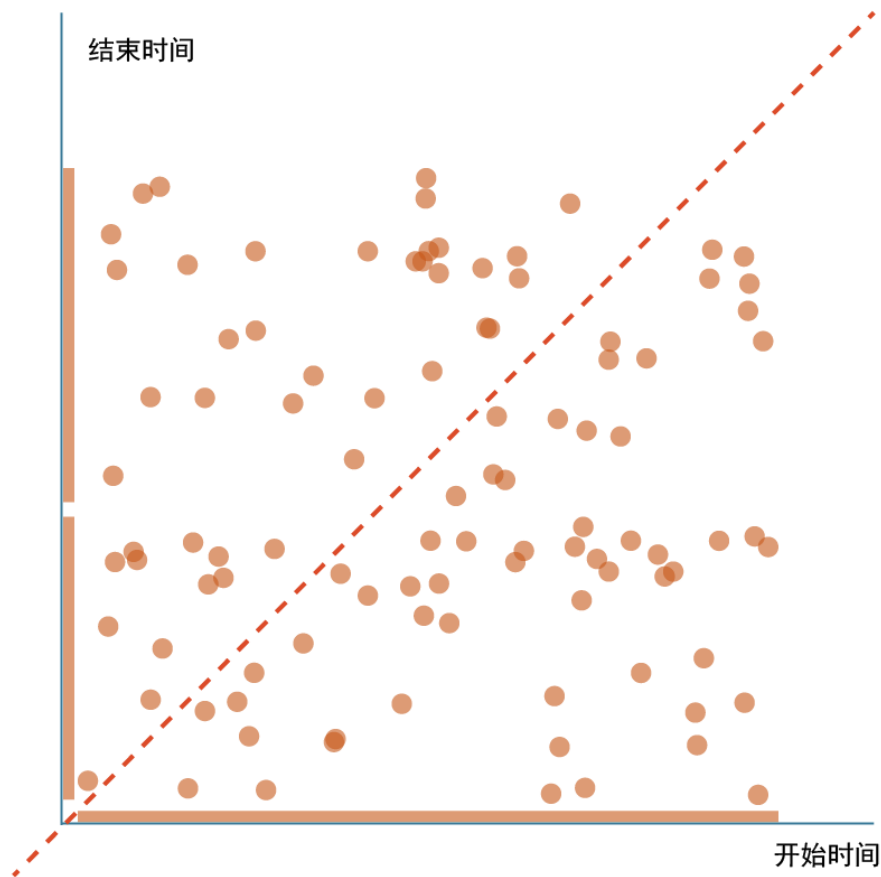
数据初探

- 确定数据属性类型
- 确定数据在各属性的值分布



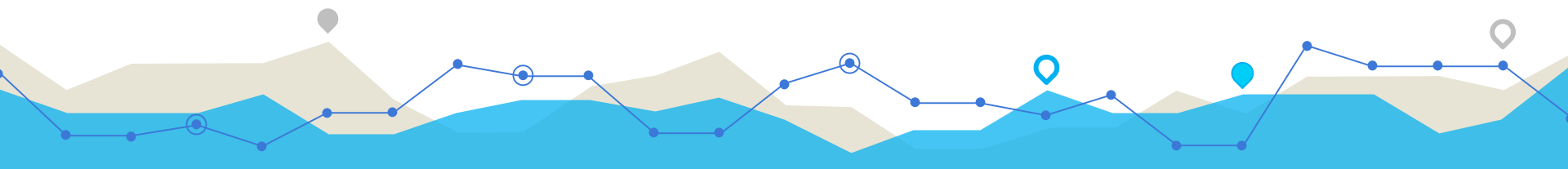
数据初探

- 确定数据属性类型
- 确定数据在各属性的值分布
- 发现不一致的数据



数据初探

- 确定数据属性类型
- 确定数据在各属性的值分布
- 发现不一致的数据
- 发现数据的误差

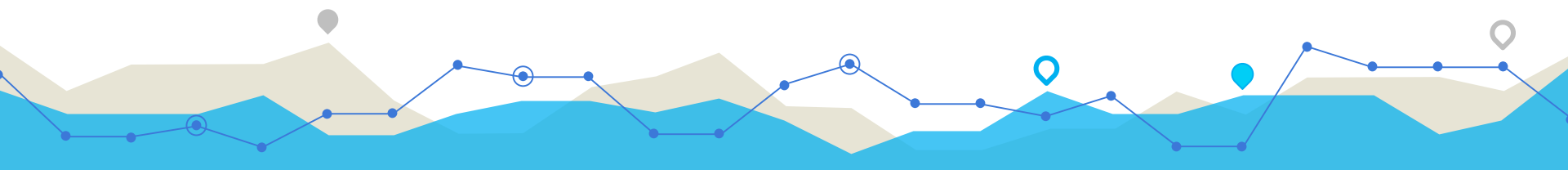


数据初探

- 确定数据属性类型
- 确定数据在各属性的值分布
- 发现不一致的数据
- 发现数据的误差

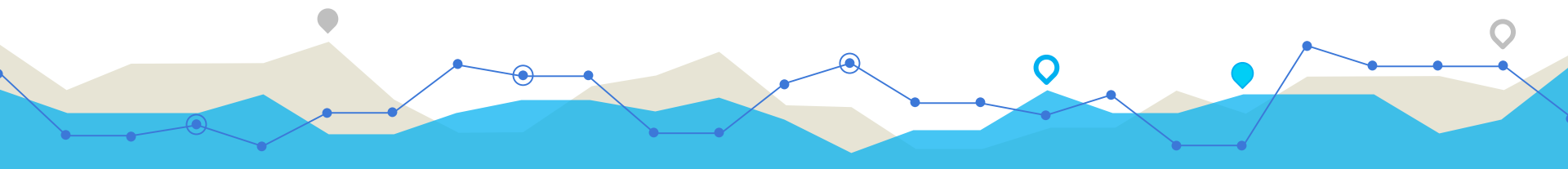
数据清洗

- 确定数据属性类型
- 确定数据在各属性的值分布
- 纠正数据的不一致性
- 校准数据



数据初探

- 确定数据属性类型
- 确定数据在各属性的值分布
- 发现不一致的数据
- 发现数据的误差
- 数据维度相关性*



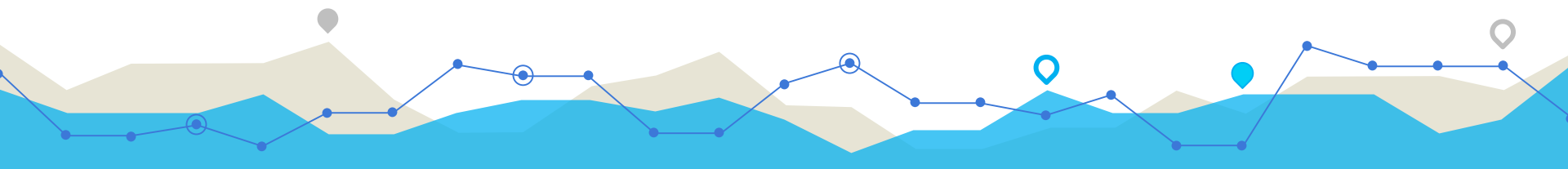
数据处理 —— 统计

方法

- 数据分布
 - 均值、方差、众数、分位数
- 回归方法
 - 线性回归
 - 逻辑回归
- 多元统计分析
 - 变量相关性

工具

- 软件类工具
 - SPSS
 -
- 脚本类工具
 - MATLAB
 - R
 - Python



数据处理 —— 降维(略)

降维的目的：

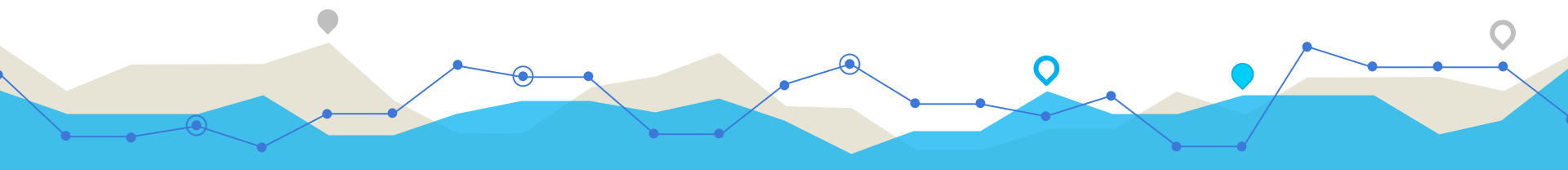
使数据在低维的距离尽量与在高维的距离保持一致

线性

- PCA、MDS、NMF、.....

非线性

- LLE、Iso-map、SOM、.....



数据处理 —— 相似度量

类别型

集合(杰卡德)
相似度

海明距离

高维数值型

曼哈顿距离
(L1范数)

欧氏距离
(L2范数)

夹角余弦

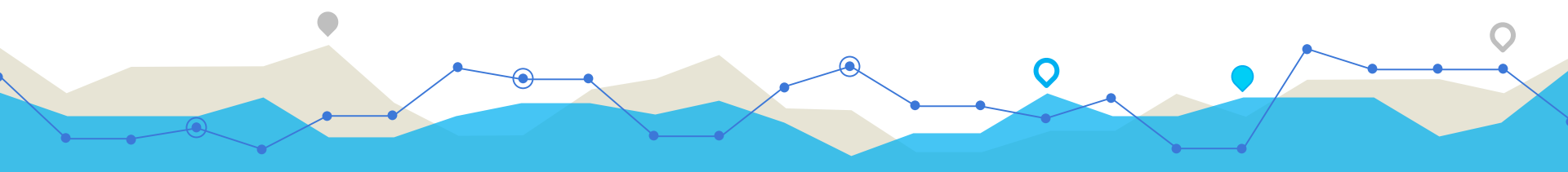
有序型

皮尔逊相关
系数

动态时间扭
曲

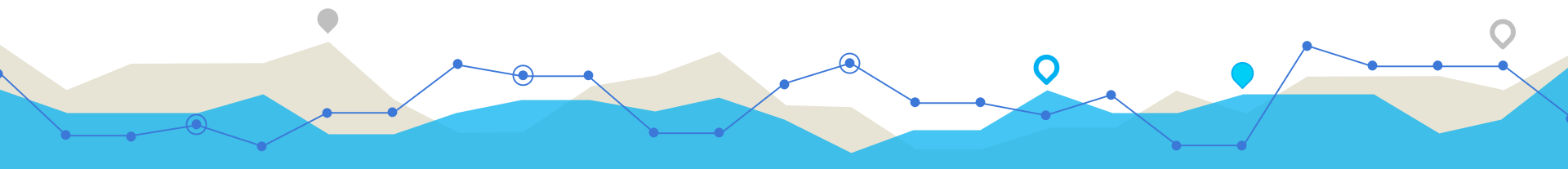
最大公共子
序列

自定义距离

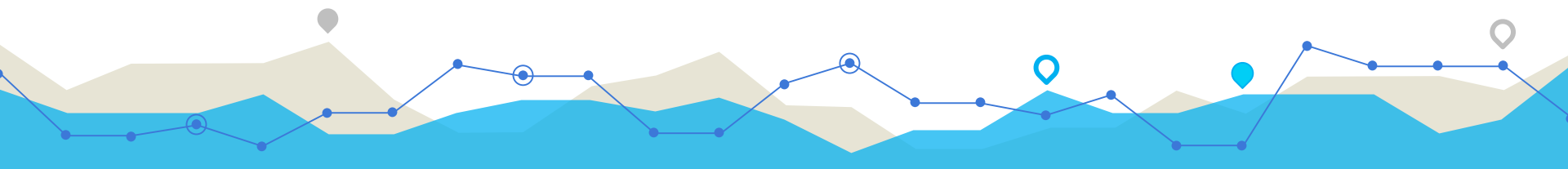
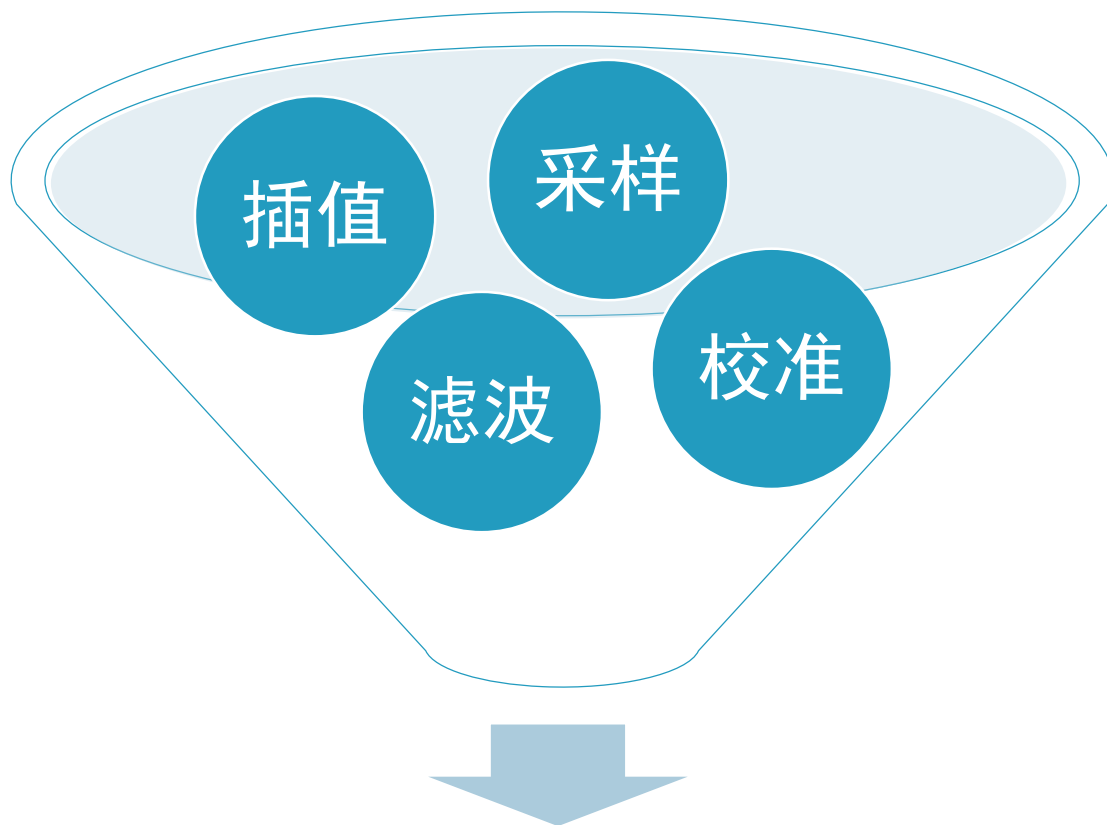


数据处理 —— 聚类

- K-均值家族
- 层次聚类
 - 自底向上
 - DBSCAN算法
 - 自顶向下
 - Graph-cut算法



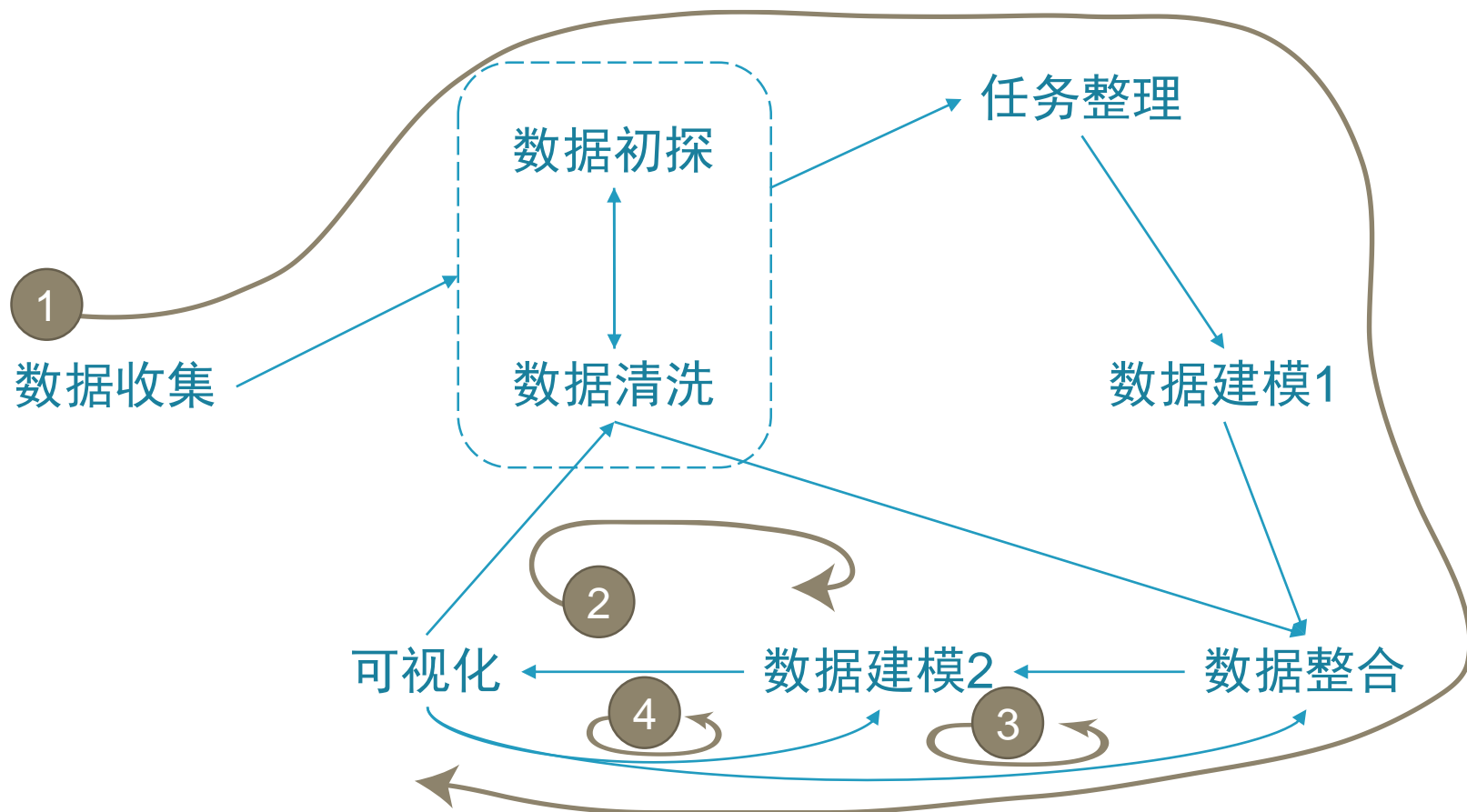
数据处理 —— 其它





应用 2

数据分析流程



完整案例 —— 维基百科热词的时序排名可视化











- 原始数据：维基百科点击日志

- 任务

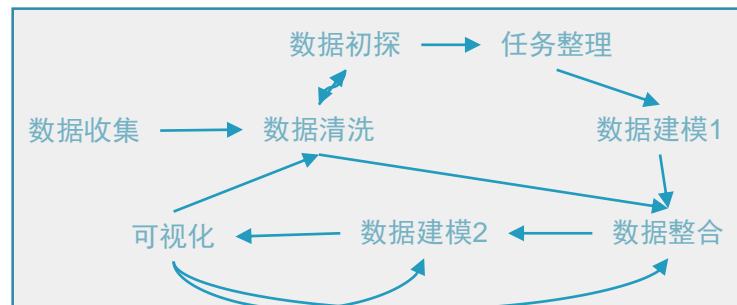
- 可视化topk页面的走势
- 分析topk页面之间的关系

- 提取数据属性

- 时间
- 排名
- 关系

Rank	Article	Class	Views	Image
1	<i>Pokémon Go</i>		4,778,652	
2	<i>Theresa May</i>		1,738,109	
3	<i>Mike Pence</i>		1,651,153	
4	<i>Sultan (2016 film)</i>		1,220,923	
5	<i>UFC 200</i>		1,139,080	

数据处理



数据收集*

- 维基百科访问日志

数据清洗

- 删除乱码数据

数据初探

- 聚合页面点击率

数据清洗

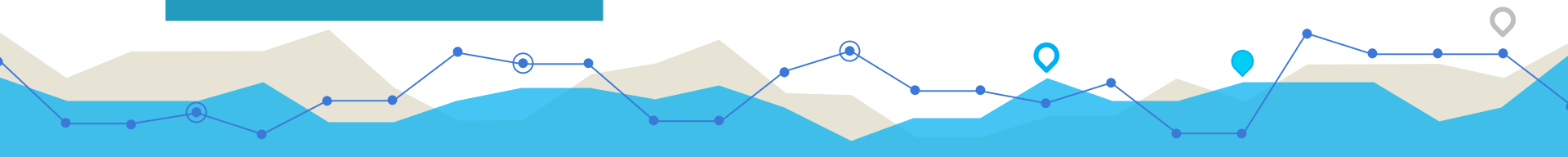
- 删除index等无效页面

任务整理

- topk页面的走势
- topk页面之间的关系

数据整合*

数据建模*



数据收集

- 维基百科访问日志

<https://dumps.wikimedia.org/other/pagecounts-raw/>

Page view statistics for Wikimedia projects

(For up-to-date information (outages, ...) about this dataset, please consult the [dataset's wiki page](#).)

Pagecount files per year

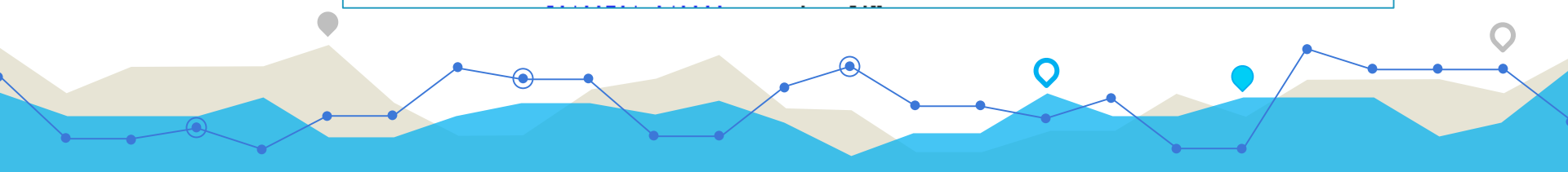
- [2007](#)
- [2008](#)
- [2009](#)
- [2010](#)
- [2011](#)
- [2012](#)
- [2013](#)
- [2014](#)
- [2015](#)
- [2016](#)

Index of page view statistics for 2016-07

Pagecount files for 2016-07

Check the [hashes](#) after your download, to make sure your files arrived intact.

- [pagecounts-20160701-000000.gz](#), size 72M
- [pagecounts-20160701-010000.gz](#), size 84M
- [pagecounts-20160701-020000.gz](#), size 90M
- [pagecounts-20160701-030000.gz](#), size 85M



数据整合

MySQL



每日top-1000页面

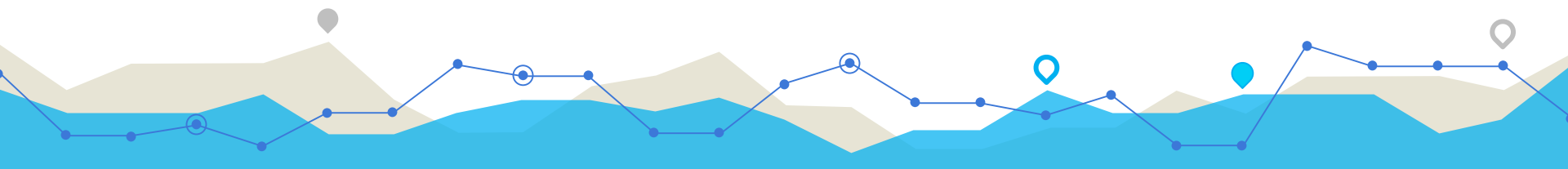
Neo4J



Pagelink关系页面

https://en.wikipedia.org/w/api.php?action=query&format=json&titles=The_Big_Bang_Theory&prop=links&pllimit=max

```
{"ns":0,"title":"Emmy Award"}  
{"ns":0,"title":"Entertainment Weekly"}  
{"ns":0,"title":"Euglossa bazinga"}  
{"ns":0,"title":"Experimental physics"}  
.....
```



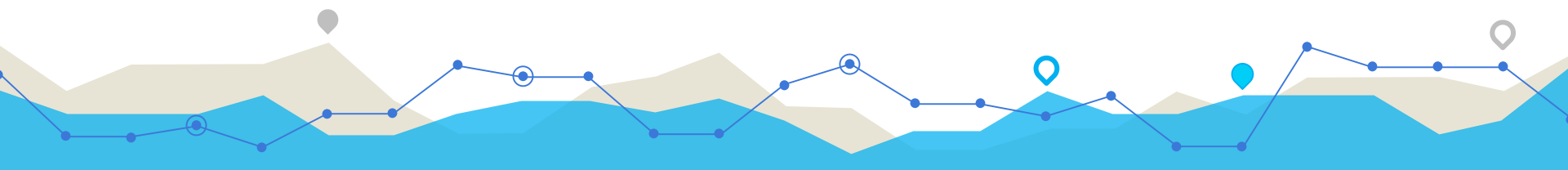
数据建模

- 某个页面的相关页面是以下页面的交集
 - 从top-1000中基于相似性查找得到的页面集合
 - 通过pagelink API找到的页面集合

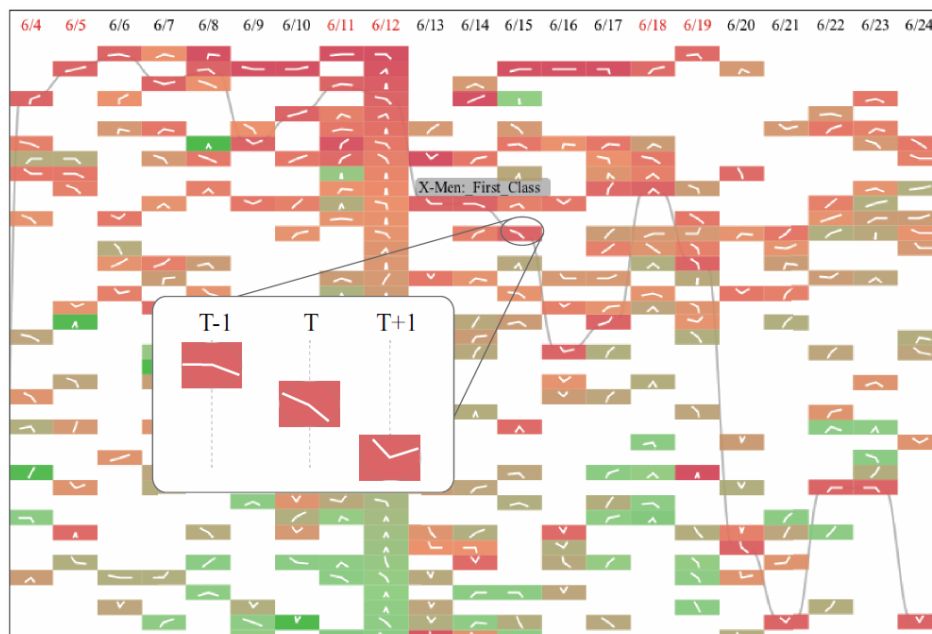
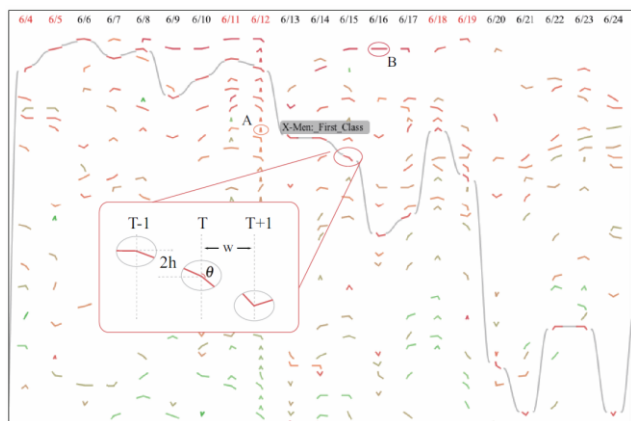
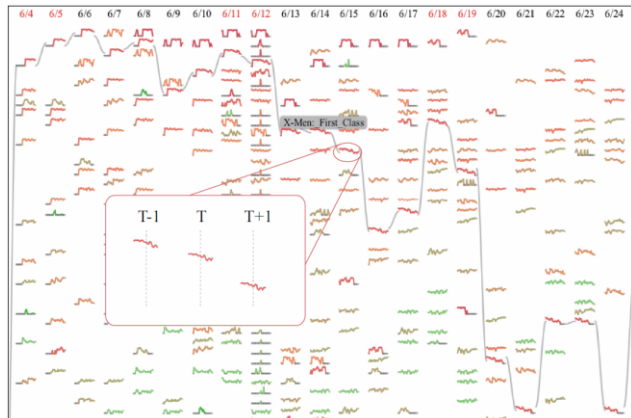
- 基于DTW的时序排名相似性建模

$$Dissim = \frac{w_{dtw} * f_{dtw} + w_{comp} * f_{comp} + w_{avgo} * f_{avgo}}{w_{dtw} + w_{comp} + w_{avgo}}$$

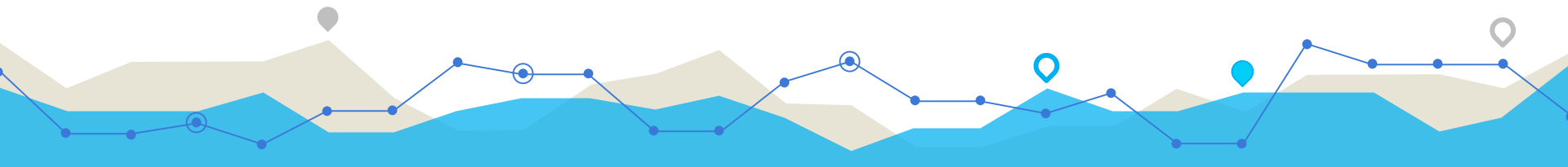
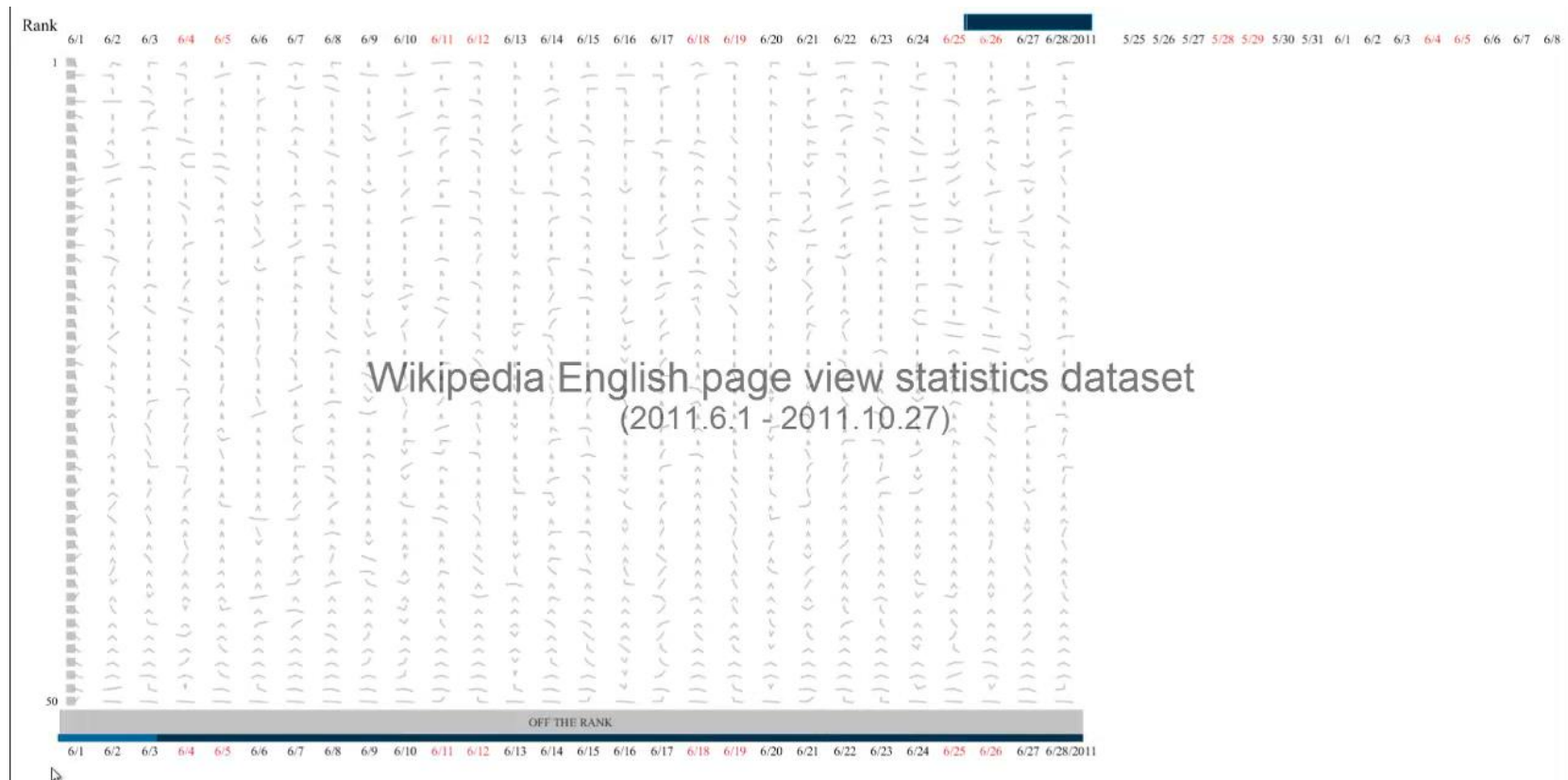
- 相似性是包括DTW因子(f_{dtw})、不连续排名损失因子(f_{comp})和平均排名因子(f_{avgo})的加权综合度量



时序排名——图元设计

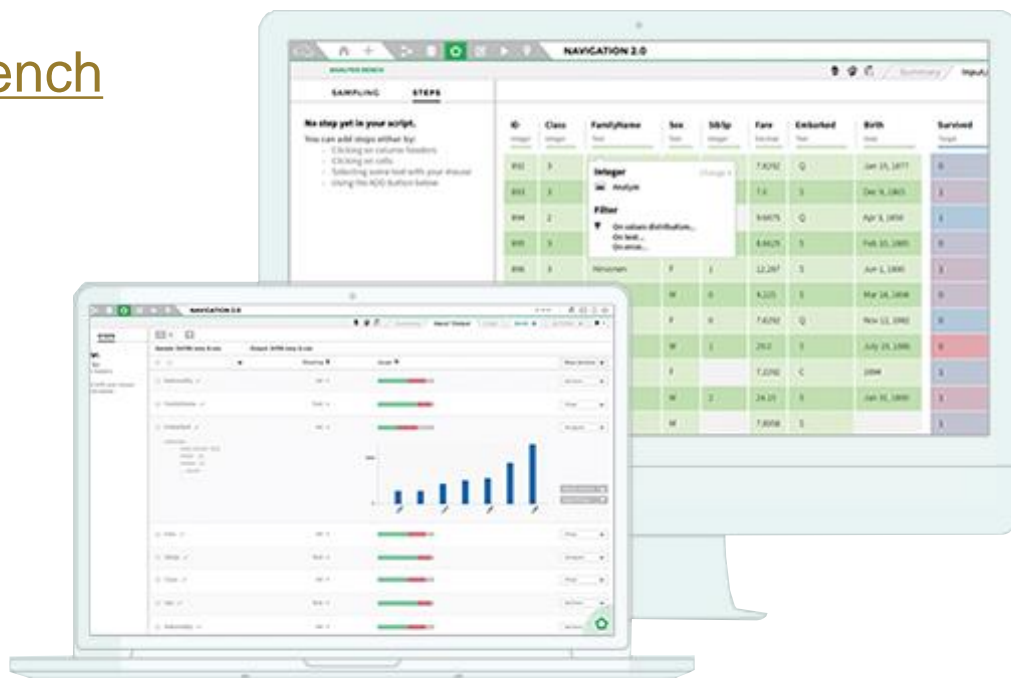


DEMO



一站式数据分析工具

- Dataiku
- DataScientistWorkbench



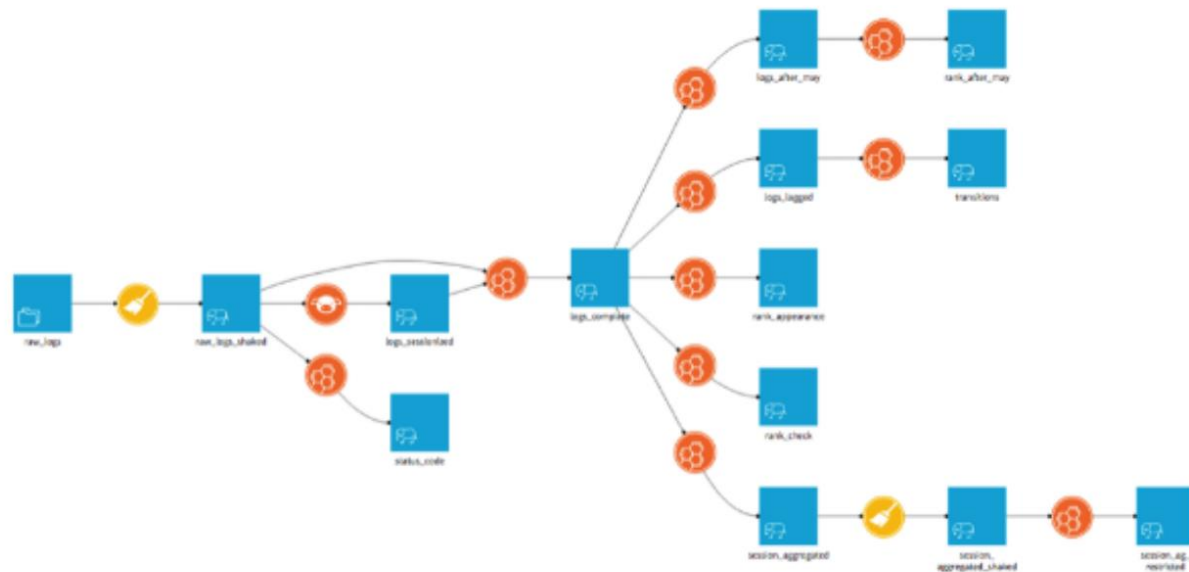
Dataiku —— 数据分析软件

• 功能

- 数据整合
- 数据清洗
- 可视化分析
- 机器学习
- 产品发布

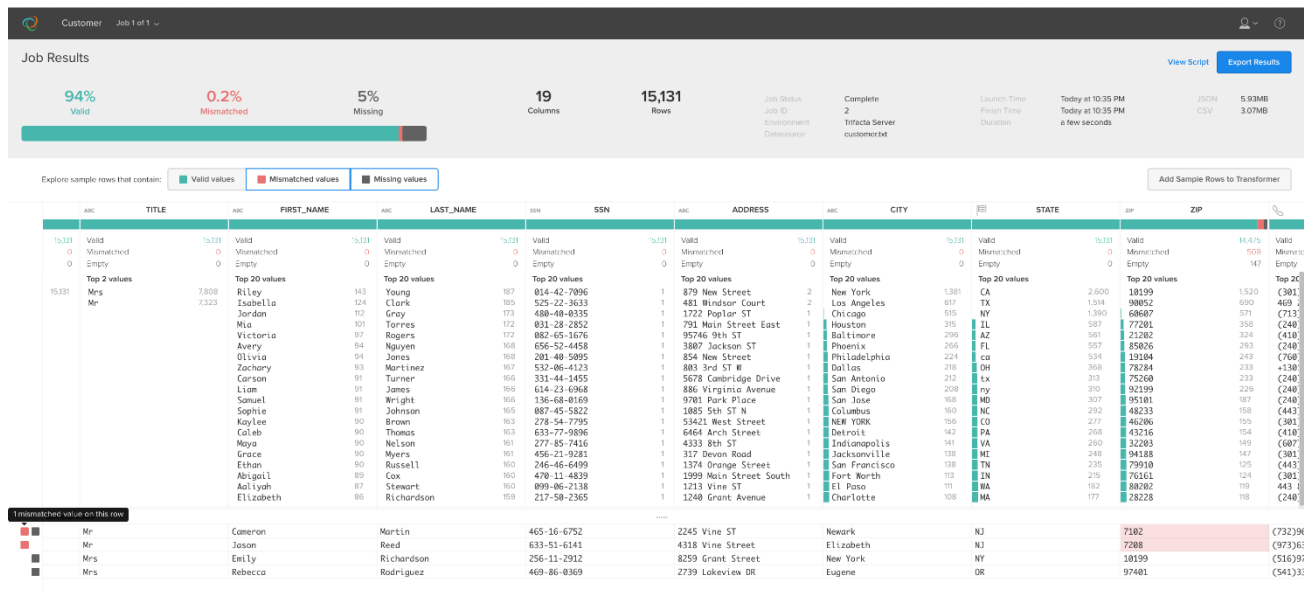
• 特点

- 功能模块化
- 界面统一

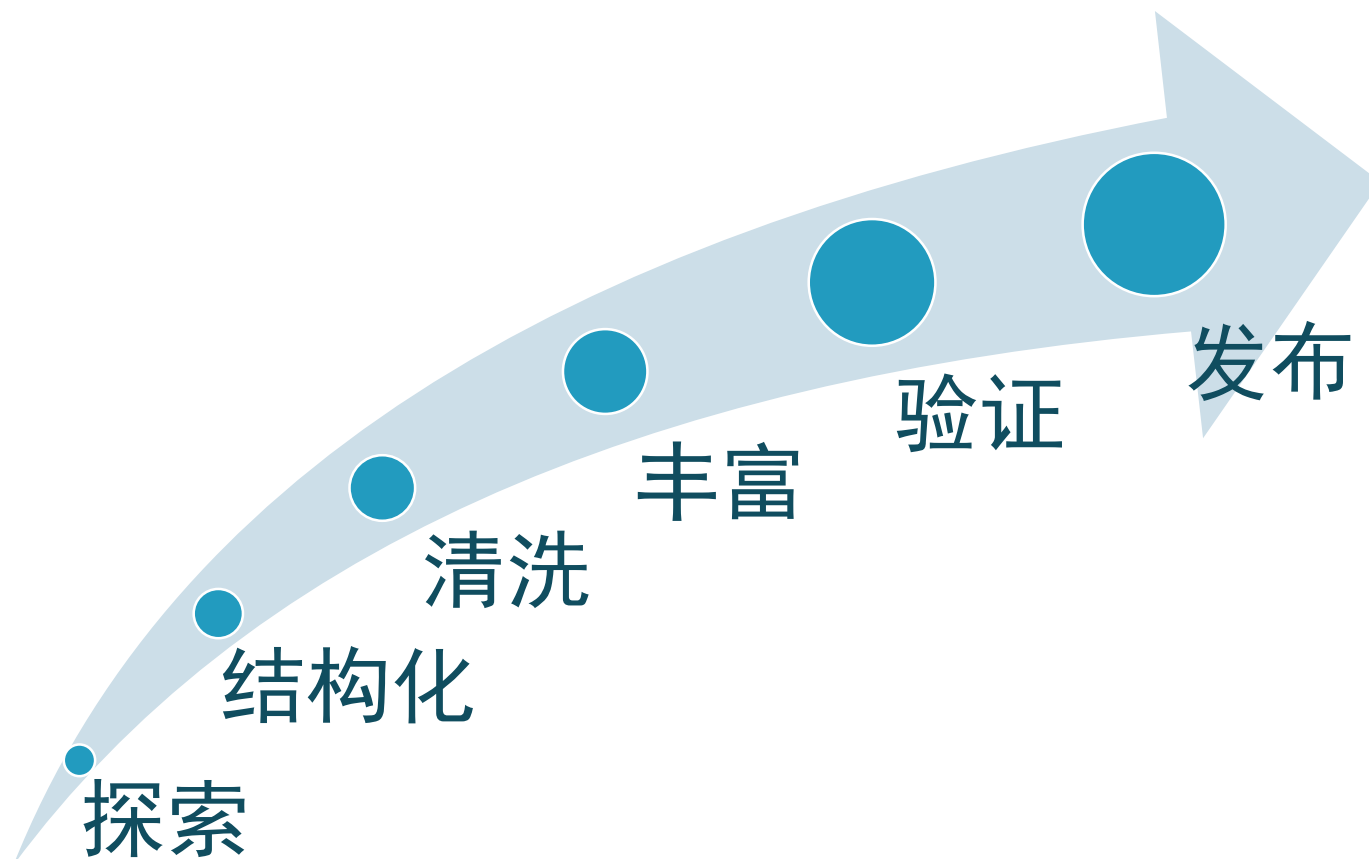


免费数据清洗工具

- Trifacta (Data Wrangler)
- OpenRefine (Google Refine)

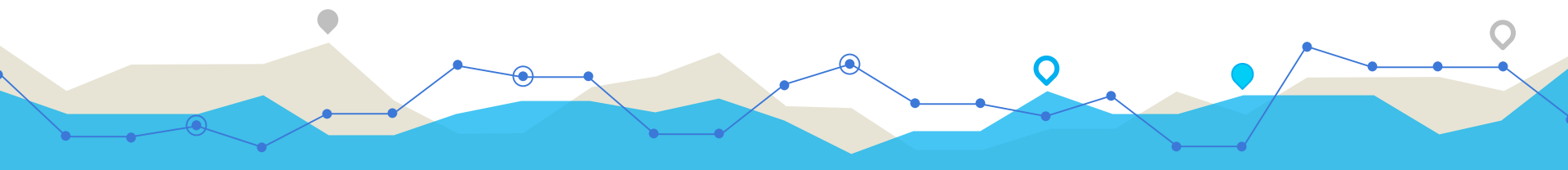


Trifacta —— 可视数据清洗工具

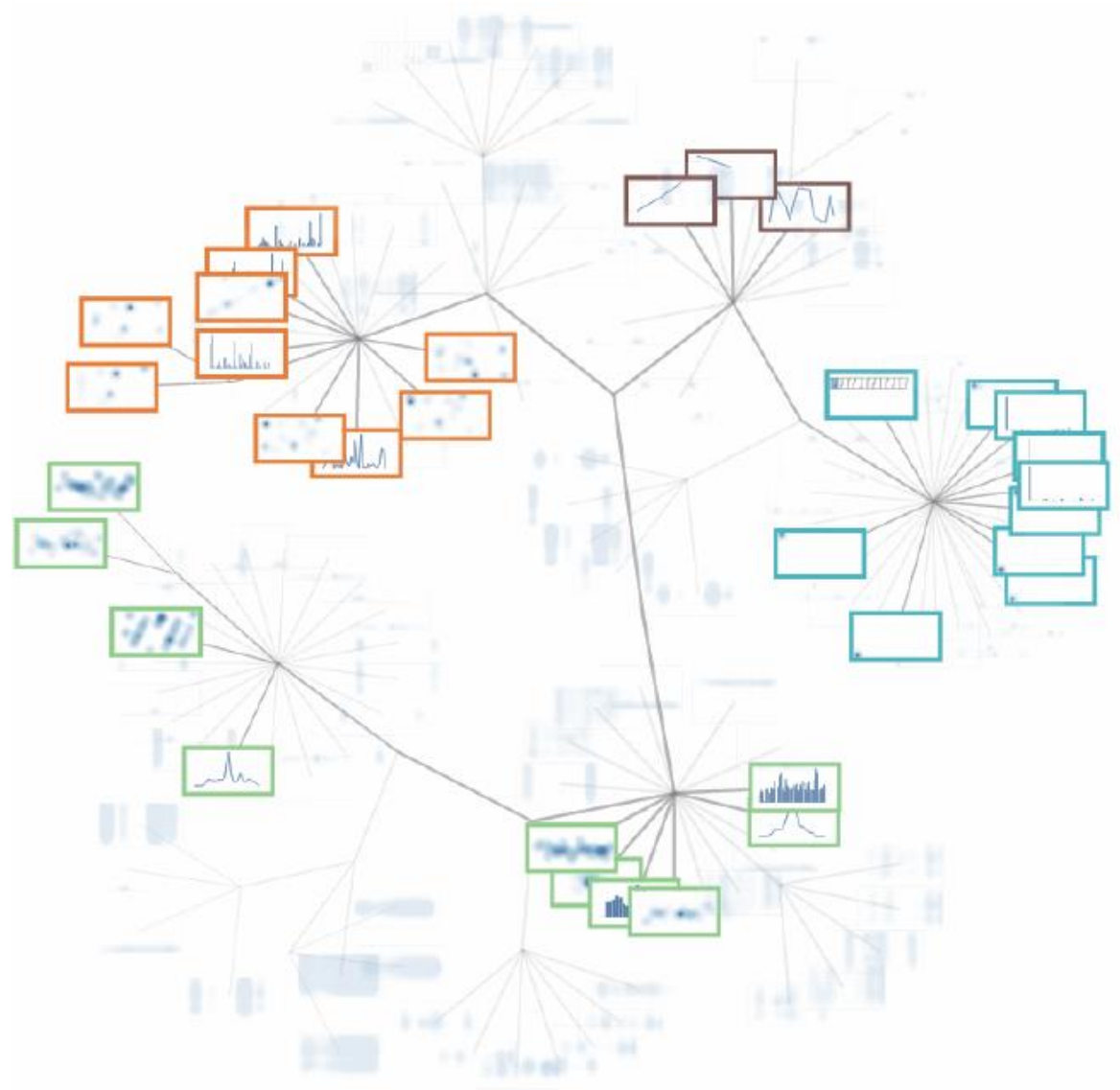


数据初探 —— DimScanner

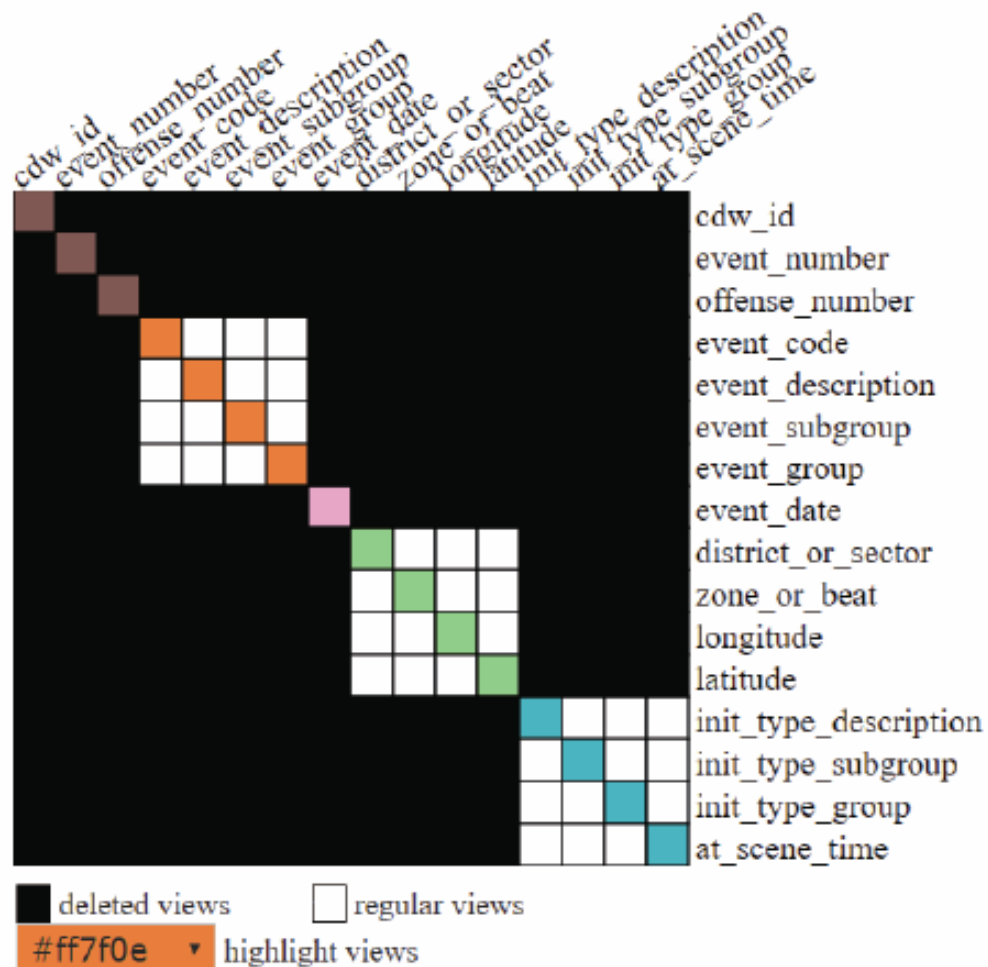
- 数据
 - 西雅图911报警数据
 - 16个维度
- 任务
 - 初探数据维度相关性



预览树

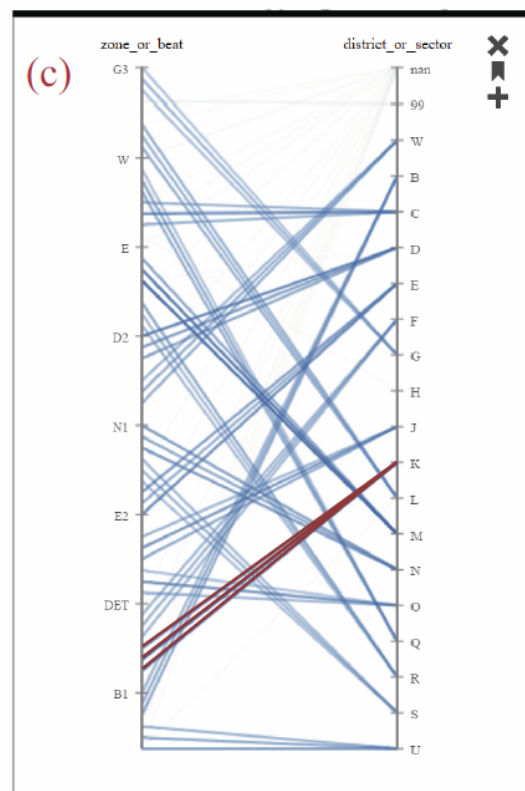
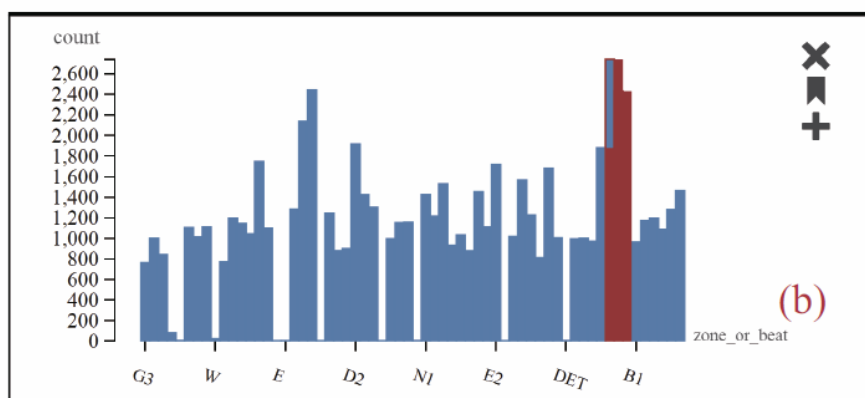
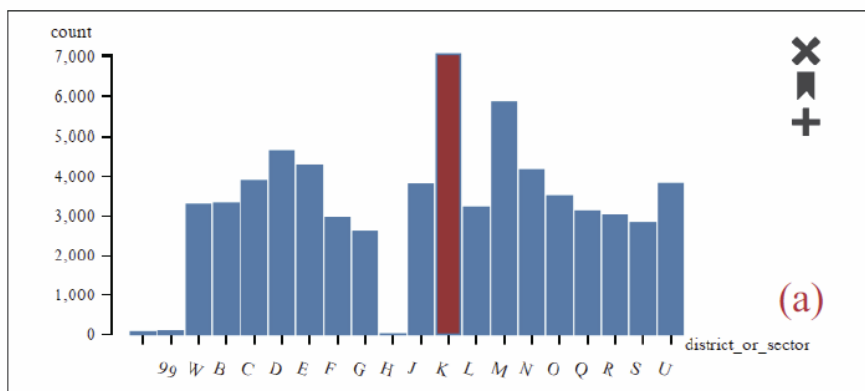


辅助交互



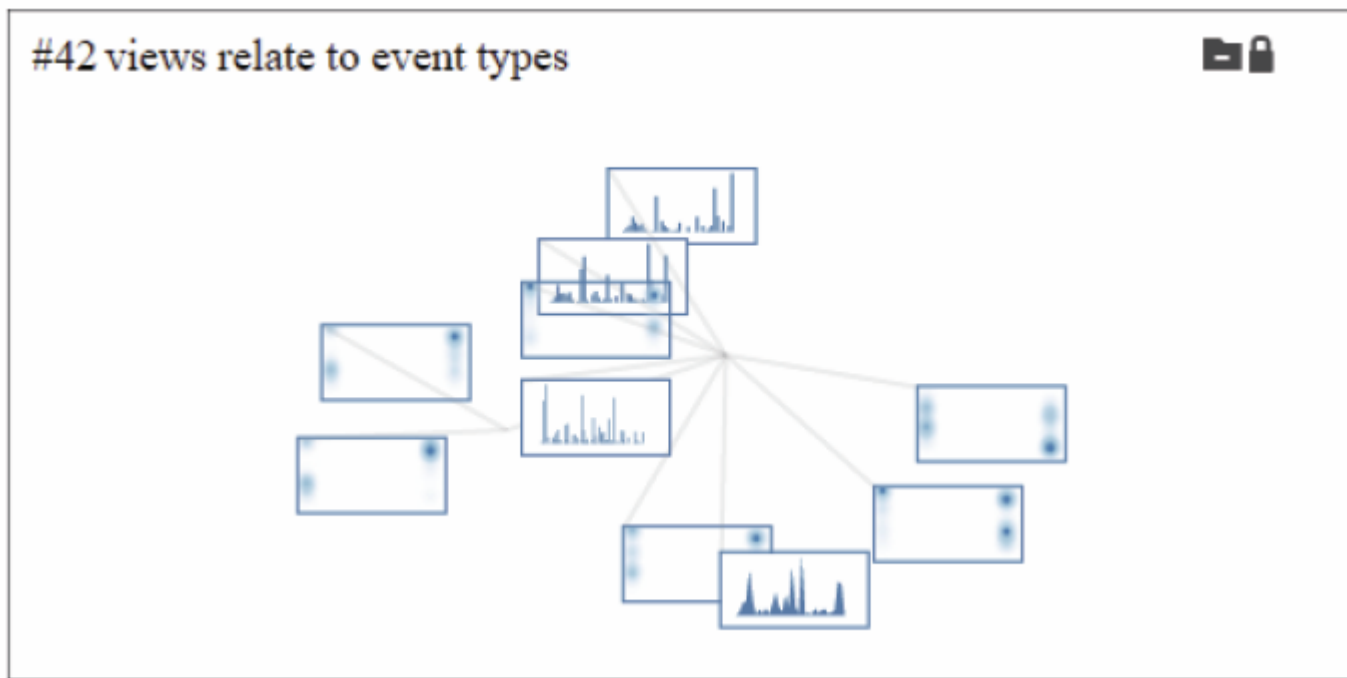
维度选择

辅助交互

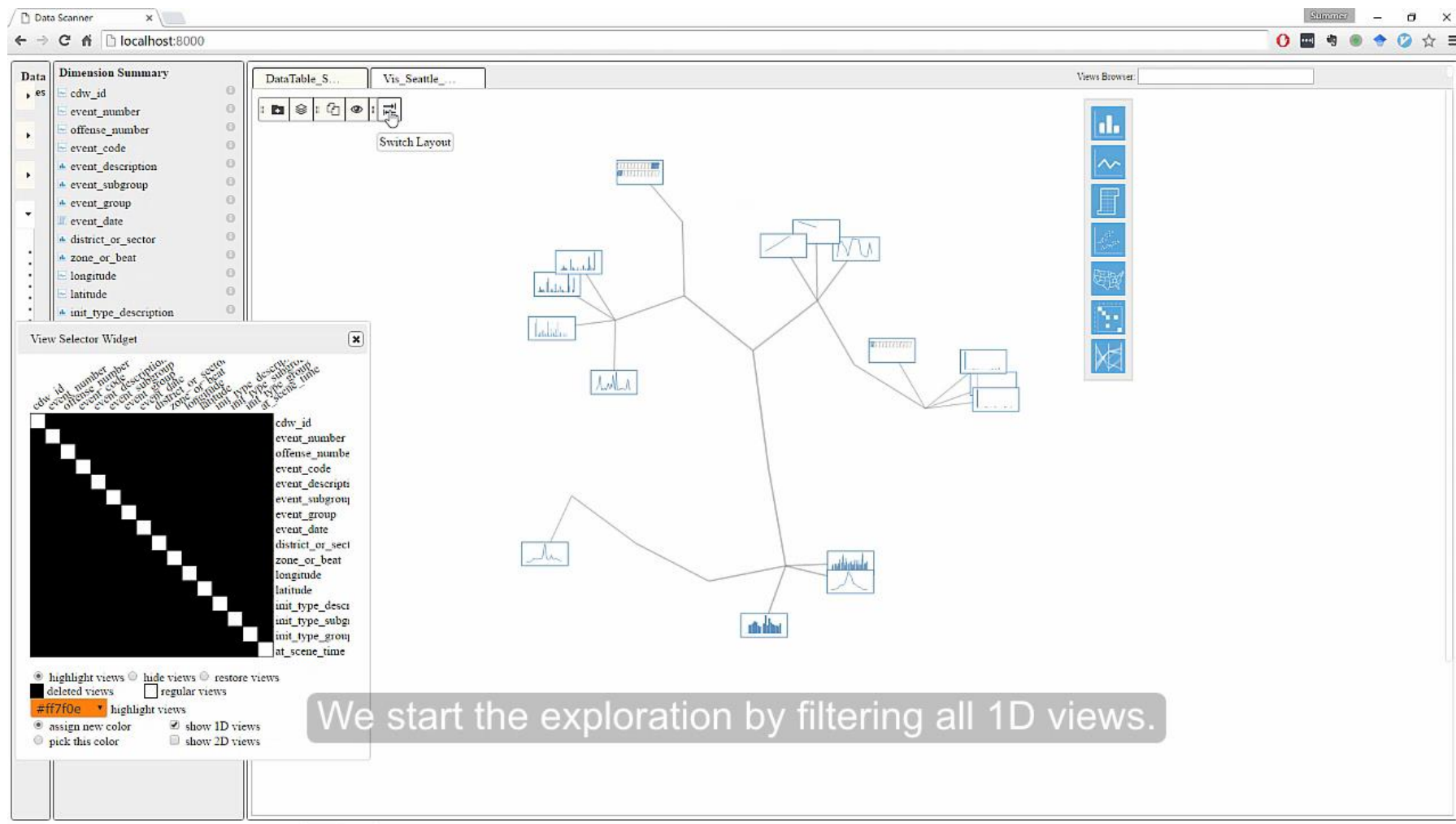


数据分布

辅助交互



数据概括



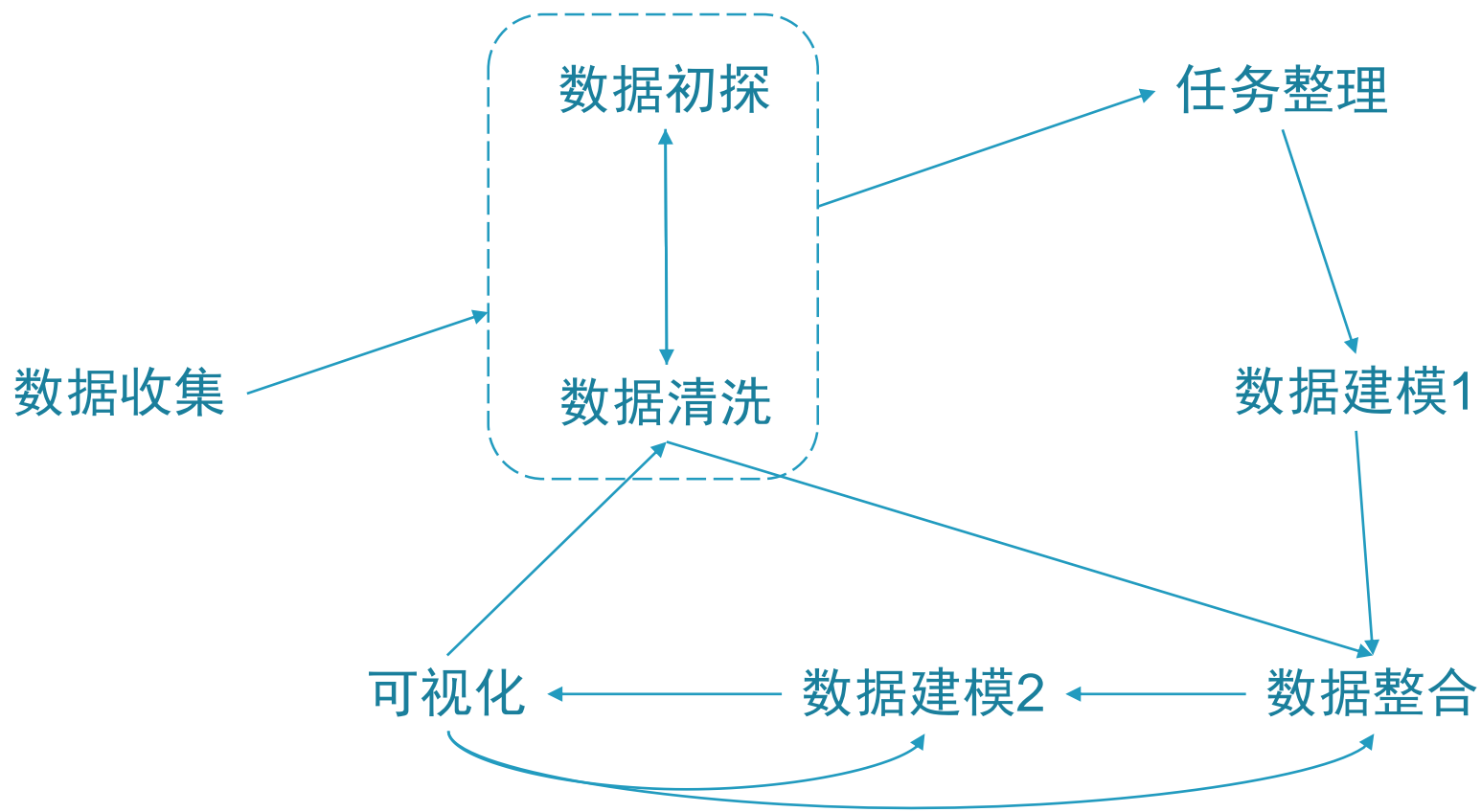


总结 3

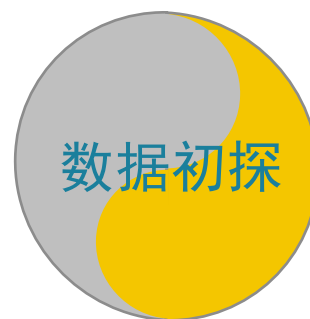
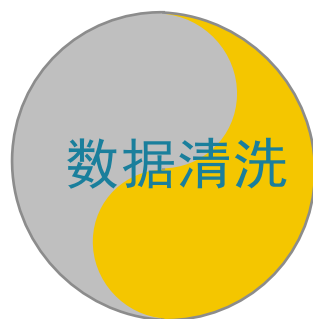
总结



总结



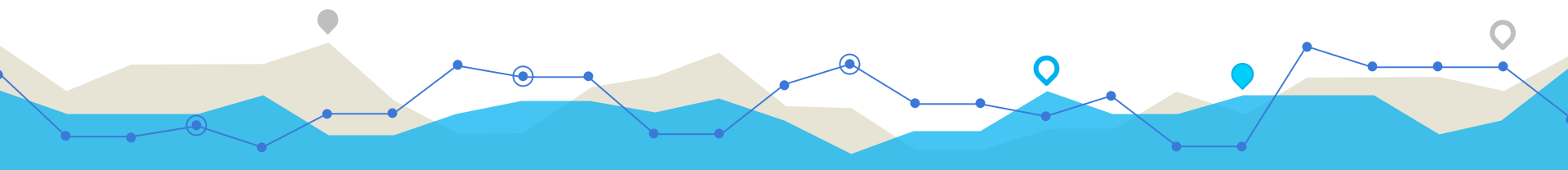
总结



总结

数据清洗和数据初探对于数据分析是至关重要的步骤

交互式可视化的辅助能够大幅度降低数据清洗和数据初探的成本



谢谢

夏菁

 summer_179279 /

jjane.summer@gmail.com

